

University of Dundee

Patterns of polymorphism and linkage disequilibrium in cultivated barley

Comadran, Jordi; Ramsay, Luke; MacKenzie, Katrin; Hayes, Patrick; Close, Timothy J.; Muehlbauer, Gary

Published in:
Theoretical and Applied Genetics

DOI:
[10.1007/s00122-010-1466-7](https://doi.org/10.1007/s00122-010-1466-7)

Publication date:
2011

Licence:
CC BY-NC

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):
Comadran, J., Ramsay, L., MacKenzie, K., Hayes, P., Close, T. J., Muehlbauer, G., Stein, N., & Waugh, R. (2011). Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theoretical and Applied Genetics*, 122(3), 523-531. <https://doi.org/10.1007/s00122-010-1466-7>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Patterns of polymorphism and linkage disequilibrium in cultivated barley

Jordi Comadran · Luke Ramsay · Katrin MacKenzie ·
Patrick Hayes · Timothy J. Close · Gary Muehlbauer ·
Nils Stein · Robbie Waugh

Received: 18 June 2010 / Accepted: 30 September 2010 / Published online: 13 November 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract We carried out a genome-wide analysis of polymorphism (4,596 SNP loci across 190 elite cultivated accessions) chosen to represent the available genetic variation in current elite North West European and North American barley germplasm. Population sub-structure, patterns of diversity and linkage disequilibrium varied considerably across the seven barley chromosomes. Genes rich and rarely recombining haplotype blocks that may represent up to 60% of the physical length of barley

chromosomes extended across the ‘genetic centromeres’. By positioning 2,132 bi-parentally mapped SNP markers with minimum allele frequencies higher than 0.10 by association mapping, 87.3% were located to within 5 cM of their original genetic map position. We show that at this current marker density genetically diverse populations of relatively small size are sufficient to fine map simple traits, providing they are not strongly stratified within the sample, fall outside the genetic centromeres and population sub-structure is effectively controlled in the analysis. Our results have important implications for association mapping, positional cloning, physical mapping and practical plant breeding in barley and other major world cereals including wheat and rye that exhibit comparable genome and genetic features.

Communicated by J. Snape.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1466-7) contains supplementary material, which is available to authorized users.

J. Comadran · L. Ramsay · R. Waugh (✉)
Genetics Programme, Scottish Crop Research Institute,
Dundee DD2 5DA, Scotland, UK
e-mail: Robbie.Waugh@scri.ac.uk

K. MacKenzie
BIOSS, Invergowrie, Dundee DD2 5DA, Scotland, UK

P. Hayes
Oregon State University, Barley Project Crop Science Bldg.
30th and Campus Way, Corvallis, OR 97333, USA

T. J. Close
Department of Botany and Plant Sciences,
University of California, Riverside, CA 92521, USA

G. Muehlbauer
Department of Agronomy and Plant Genetics,
University of Minnesota, St. Paul, MN 55108-6026, USA

N. Stein
Leibniz Institute of Plant Genetics and Crop Plant
Research (IPK), Corrensstr. 3, 06466 Gatersleben, Germany

Introduction

Genome-wide association studies are currently commonplace in the search for genes controlling human genetic disorders (Kruglyak 2008) but are in their infancy in plant genetic research, particularly in crops (Zamir 2008). This is despite the approach sparking considerable interest and having tremendous potential for identifying the genes underlying agriculturally important traits, particularly through exploiting the tremendous wealth of highly replicated historic phenotypic information that is potentially available to researchers through national plant trialing and registration schemes (Waugh et al. 2009). The exceptions are the model plants with fully sequenced genomes (The Arabidopsis Initiative 2000; Goff et al. 2002; Jun Yu et al. 2002) and maize in both private and public sectors (Chandler and Brendel 2002; Lawrence et al. 2008) where good progress has recently been made.

In the majority of crops, while interest in association mapping is considerable, implementation has so far been restricted. A common reason appears to be the lack of high-plex, low cost, robust and informative marker platforms that facilitate sufficiently dense genome-wide analysis of molecular polymorphism to assist subsequent association analysis (Zhu et al. 2008). Furthermore, much of the theory and practice of association mapping has been established in heterozygous outbreeding species (Ersoz et al. 2009), and there is limited detailed information on both the extent and patterns of polymorphism, and their consequent impact on genome-wide association scans, specifically in autogamous crop populations where linkage disequilibrium (LD) is predicted to be extensive (Caldwell et al. 2006).

Barley is a diploid inbreeding crop plant that, over the last 10,000 years (Badr et al. 2000), has gone through complex and rapid evolution imposed by the dual bottlenecks of domestication and breeding. While it is a fairly strict inbreeder, since the late 19th century, forced cross-pollination imposed by breeders, followed either by several rounds of inbreeding or, in the second half of the 20th century, by the generation of F1 doubled haploid's coupled with heavy selection for desirable phenotypes, has generated a unique population of homozygous 'elite' inbred cultivars with intricate and complex pedigrees. We recently proposed (Rostoks et al. 2006) that this pseudo-outbred 'elite genepool' could be used effectively for medium resolution genome-wide association scans for trait-gene identification using a manageable number (100–1,000's) of robust bi-allelic markers. From an applied standpoint this is particularly attractive for trait dissection as the elite genepool contains the majority of the genetic variation currently being manipulated by breeders in contemporary crop improvement schemes. Diagnostics for positive alleles for relevant traits in this germplasm will help to facilitate the realization of 'predictive breeding' often discussed as necessary to meet the world's future demands for food and feed (FAO 2002). Understanding the patterns of polymorphism in this type of material and the possible limitations it has for genome-wide association scans is therefore an important step that will have significant fundamental and translational outcomes.

We assembled 190 elite cultivated accessions from two large association genetics programs in the US (BarleyCAP, <http://www.barleycap.org/>) and UK (AGOUEB, <http://www.agoueb.org/>) chosen to represent the available genetic variation in current elite North West European and North American barley germplasm. The sample encompasses elite lines from three major biotypes present in barley germplasm (Malysheva-Otto et al. 2006) that we genotyped with 4,596 SNP loci. We used this genotypic information to carry out a genome-wide analysis of population substructure, investigate patterns of diversity and LD and

interpret the implications of our findings for association mapping. Ultimately, we demonstrated the validity of genome-wide association mapping in this germplasm by fine-mapping each of 2,132 SNP markers that exhibited minimum allele frequencies (MAFs) higher than 0.10. The possibilities for identifying and validating candidate genes that underlie positive marker-trait association are discussed.

Materials and methods

Plant materials

We assembled a total of 190 elite cultivated accessions from two large association genetics programs in the US (BarleyCAP, <http://www.barleycap.org/>) and UK (AGOUEB, <http://www.agoueb.org/>) chosen to represent the available genetic variation in current elite North West European and North American barley germplasm (Table S1).

Genotyping

4,596 high confidence SNPs were assayed across the association mapping panel. These SNPs were incorporated into three IlluminaTM GoldenGate Pilot Oligo Pool Assays (POPA 1, 2, and 3) as described by Close et al. (2009). In the current experiment, 671 SNP assays were considered as 'failures' and omitted from the dataset. Of the remaining 3,925 SNPs, 2,943 had previously been incorporated into a combined genetic map, and 982 were unmapped. Ambiguous calls were coded as 'missing data' in all analyses. A sub-set of 2,709 mapped SNPs with 'missing data' <10% was collected for all 190 elite barley accessions (Table S1). Of these, 2,132 had MAF > 10% providing data matrices of 2,943 × 190 and 2,132 × 190 loci which were used to explore patterns of genetic diversity, population structure and linkage disequilibrium, and genome-wide association scans (GWASs), respectively. All genotyping assays were conducted at the Southern California Genotyping Consortium at the University of California, Los Angeles.

Population structure and patterns of diversity

We calculated a phylogenetic tree using the neighbour-joining (NJ) tree building and clustering algorithm implemented in the PHYLIP package (Felsenstein 1997). The resulting dendrogram was rooted using the *Hordeum spontaneum* line "Mehola". Principal coordinate (PCO) analysis based on simple matching of SNP alleles was performed with Genstat 11 (Payne et al. 2008). Thirdly, Bayesian clustering, again using simple matching, was applied to identify clusters of genetically similar

individuals using STRUCTURE 2.1 considering admixture (Pritchard et al. 2000b; Pritchard and Donnelly 2001), with population differentiation measured using the F_{st} estimator implemented in the STRUCTURE software.

Genome structure and linkage disequilibrium

Pair-wise measures of LD (r^2) were calculated for the selected 2,132 SNPs for each chromosome using Haploview 4.01 (Barrett et al. 2005). Only markers with $MAF > 0.1$ and pair-wise comparisons with $p > 0.001$ were considered. r^2 values were plotted as a function of genetic distance for each chromosome. Haploview was used to generate r^2 LD heat-map charts for each chromosome.

Genome-wide association

To determine markers associated with a trait of interest, SNP data were modeled using a generalized linear mixed model so that random population structure estimates could be fitted to reduce type I errors. Mapped SNP marker scores were all considered as binomial traits and only SNP markers with $MAF > 0.1$ and ‘missing data’ $< 10\%$ were used.

Mixed model methodology

We derived a relative kinship matrix (K) on the basis of simple matching coefficients from a set of random SNP data using Genstat software (Payne et al. 2008). Markers were fitted as fixed effects. Genotype was fitted as a random effect which is assumed to be distributed as $N(0, 2K\sigma_g^2)$ where K is the kinship matrix. $-\log_{10}(p \text{ value})$ scores were used as a measure of LD. For the STRUCTURE model, the resulting STRUCTURE output matrix (Q) for $k = 7$ was directly used as co-factor in the random term of the mixed model.

Results

Population structure and substructure

We genotyped 190 accessions chosen to represent the available diversity within the elite cultivated genepool from NW Europe and the USA, including a small number of foundation genotypes and key cultivars that have featured strongly in the development of contemporary barley cultivars in these regions, with barley POPA 1, 2 and 3. The assembled dataset contained a total of 1,746,480 allele assignments ordered along each barley chromosome. After manual supervision and correction, 347,118 data points,

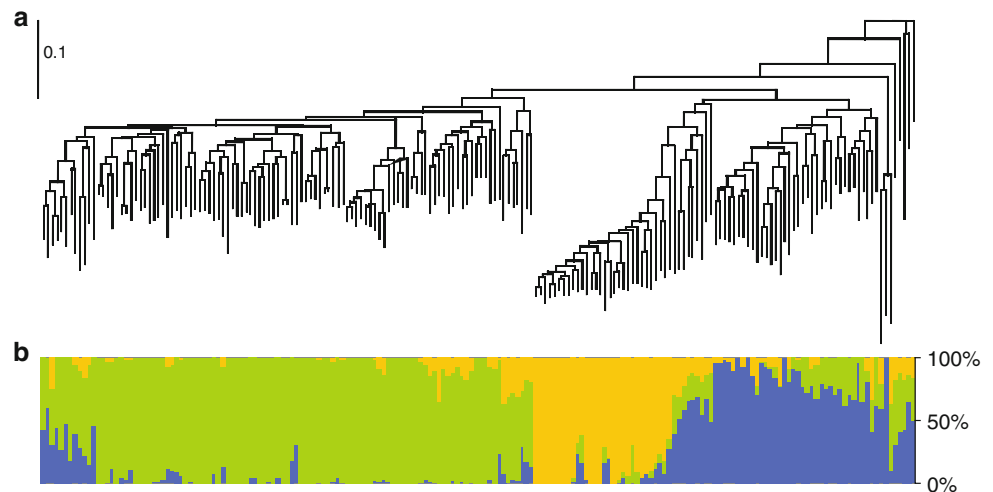
including all data from poor quality SNPs were removed from the dataset (254,980) or coded as missing (92,138) in all subsequent analyses.

We used several previously applied approaches to partition the germplasm into sub-populations based on the collected molecular polymorphism data. PCO analysis largely separated the material into well-established biotypes (Malysheva-Otto et al. 2006). PCO 1 accounted for 19.05% of the genetic variance and separated accessions according to the number of rows of seed on the mature inflorescence (2 vs. 6 rows). PCO 2, accounting for 8.72% of the variation, separated winter sown from spring sown genotypes. Three germplasm groups were observed based on the first two PCO's. The six-row spring barley accessions formed an exclusive independent cluster (Figure S1). This sub-population exhibited a fixation index, F_{st} value of 0.7903, indicative of considerable genetic differentiation from the remaining accessions (Hudson et al. 1992). They represent genetic material derived from founder lines imported into the USA from Manchuria and neighboring regions in North Eastern China that gained favor in the northern Great Plains because of their good malting quality and regional growing performance. Genotypes in the other two sub-groups will most likely have been derived from landraces originating from the ‘Fertile Crescent’ of Israel, Syria, Jordan and Iran and reached Europe and the US through well-established domestication routes (Badr et al. 2000). Population structure was also determined using the Bayesian approach implemented in the program STRUCTURE (Pritchard et al. 2000a) and the results compared (Fig. 1). STRUCTURE indicates an optimal number of groups (k) of 7 (Figure S2). However, three major groupings mirroring those observed using PCO analysis were observed at $k = 3$ and made both geographical and genetical sense. The extra groups observed using $k = 7$ (Figure S3) represent small highly differentiated germplasm sets with a narrow genetic base within the winter and two-row spring genetic backgrounds and exhibited F_{st} values of around 0.8 compared to 0.45 within in their associated major groups.

Patterns of polymorphism along barley linkage groups

Patterns of polymorphism along barley linkage groups (Fig. 2) were investigated for the three major groupings established by PCO analysis ($n = 105, 51$, and 34 for two-row spring barleys, winter barleys and six-row spring barleys, respectively). Each group exhibited a contrasting pattern of diversity that in some cases most likely reflects the selection of loci for key traits during domestication and breeding. Overall genetic diversity across the 190 lines is high and generally stable all across the genome (Fig. 2, black line). Strikingly, a clear depletion of genetic diversity

Fig. 1 Population stratification.
a Tree rooted with wild barley accession Mehola.
b STRUCTURE output for $k = 3$ mirror the three main branches of the tree and PCO groupings (Figure S2)



can be observed for all three germplasm groups on the short arm of chromosome 3H, where 11 contiguous SNP markers delimit a 2.9 cM interval that has been fixed within the cultivated germplasm examined. Linkage drag around this locus affects up to 2% of the genetic map.

Linkage disequilibrium

We explored whole genome patterns of LD using classic LD algorithms (r^2 and D') and using a mixed-model approach with population structure estimators as co-factors to account for most of the population structure effects on long-range LD. We considered it important to remove long-range LD effects because they may obscure our interpretation of both genome coverage and mapping resolution.

Classic LD algorithms

Inspection of the 2,943 genetically mapped SNPs scored across the germplasm set revealed a subset of 2,132 that exhibited a minor allele frequency (MAF) higher than 0.10 in this germplasm set with less than 0.10 of missing data. These were used in subsequent analyzes. Plotting LD as a function of genetic distance revealed extensive intra-chromosomal LD along each barley chromosome (Figure S4). Heatmap charts of the distribution of intra-chromosomal r^2 values across each barley chromosome highlight the extended LD values across the genetic centromeres (Figure S5). High LD extends outwards from these regions along the spine of each chromosome forming an axis of blocks of short-range LD. A background of long-range LD, that commonly results from population sub-structure and admixture within a germplasm set (Ersoz et al. 2008), was observed for all seven chromosomes (Figure S5).

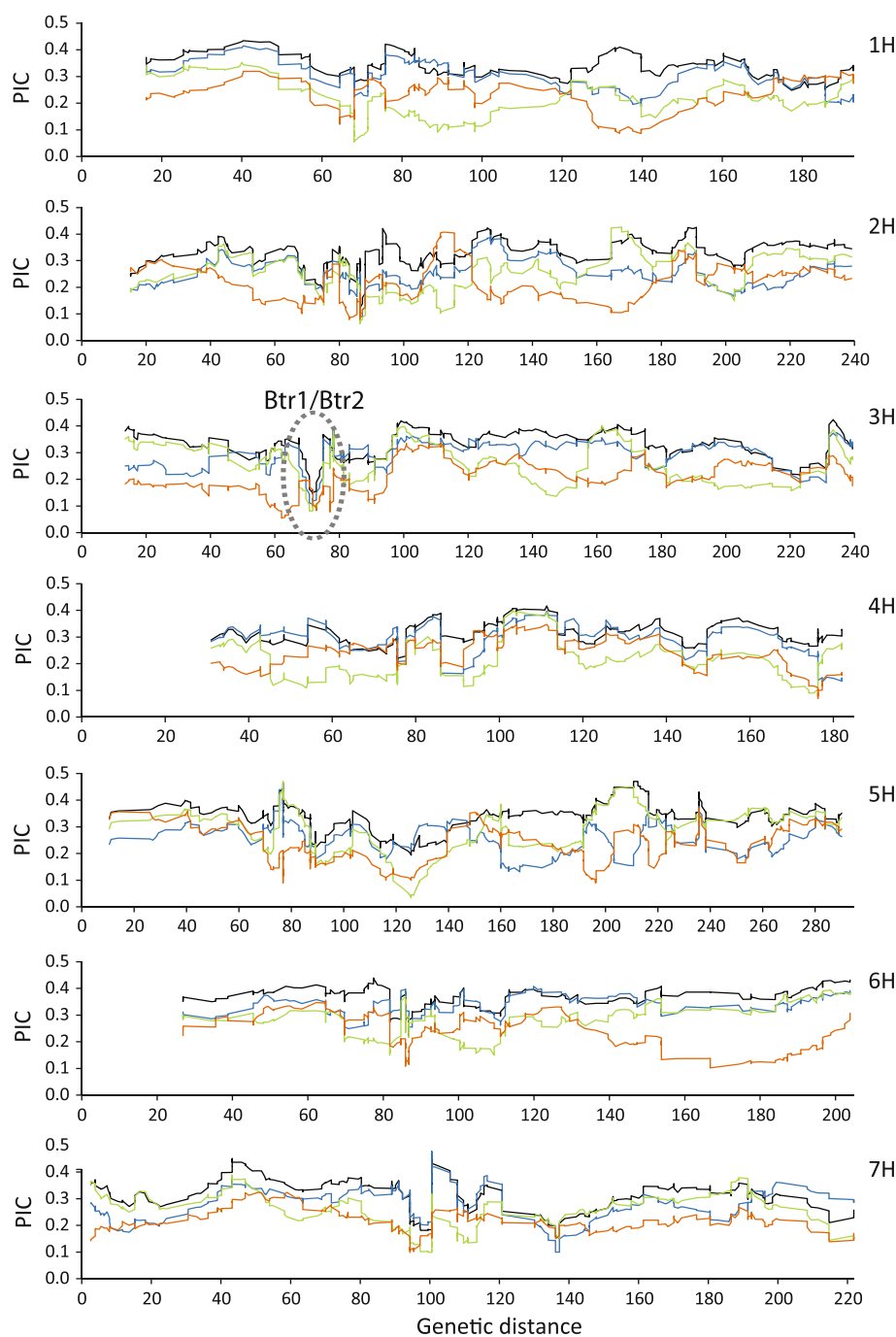
Mixed model

2,132 mapped SNP loci were used for a $2,132 \times 2,132$ genome-wide association (GWA) scan using a mixed model. In this analysis each SNP at a time is removed from the marker dataset and used as a simple trait to be mapped by the remaining 2,131 SNP markers. A heatmap chart of the distribution of pairwise $-\log_{10}(p \text{ value})$ scores of the $2,132 \times 2,132$ SNP markers was then built (Fig. 3). High LD extends along the spine of each chromosome, and the background of long-range LD is drastically reduced compared to that observed using classical r^2 and D' algorithms (Text S1; Figures S6, S7).

Mapping resolution

As a result of the previous exercise we now had two map positions for each SNP: the original genetic map position based on the consensus map of Close et al. (2009) and the position where the SNP has been mapped by association mapping. The genetic distance between the most significantly associated marker from the GWAS and the map location of the mapped SNP on the bi-parental consensus map can then be used to evaluate the mapping resolution attained and the amount of putative false positives. Figure 4 summarizes the results: 50% of the 2,132 SNP markers map within 1 cM of their original genetic map position with a fast decay of LD following genetic distance until 91.2% map within 10 cM (87.3% of the SNP markers map within 5 cM). A sub-set of 126 SNP markers mostly located in centromeric regions, had identical genotypic profiles to at least another SNP marker. Due to a lack of recombination, most of the centromeric markers mapped with co-segregating markers within the same 'genetic centromere', despite the fact that they may be separated by

Fig. 2 Diversity across barley genome. Polymorphic information content (PIC) was calculated using the method of Botstein et al. (1980). PIC was averaged across a sliding window of 20 adjacent loci with a step of one and plotted against the linkage map. STRUCTURE groupings at $k = 3$ were used to detect and remove clear cases of admixture between the three main groups so the patterns of polymorphism do reflect history of breeding and selection within them. Manchurian types sampled are highly inbred with reduced effective sample size, thus further investigation will be required. *black* whole sample, *blue* winter barleys, *green* two-row spring barleys, *orange* six-row spring barleys (Manchurian types). *Btr1/Btr2* Diversity depletion. The strong signature of selection observed co-locates with the map position of the brittle rachis trait



very large physical distances. Most of the SNP markers mapping between 5 and 10 cM had low significance values and fell into regions of the original consensus map with the lowest SNP density.

Conservation of synteny among grass genomes can potentially provide a more accurate estimation of resolution based on the gene content of the comparative genetic interval in fully sequenced models. After removing all pairwise SNP markers with inter-chromosomal associations, those with poor BLAST hits to the rice genome sequence

and those where barley-rice synteny was not conserved, we selected 685 SNPs in genes that we were confident could reasonably define putative gene content in the intervening regions in the rice genome. We then counted the number of gene models in rice that separated each ‘test’ SNP from its most strongly associated SNP that we identified by GWAS. Figure 5 summarizes the results: 50% of the genes containing the ‘test’ SNPs were located within 27 rice gene models from the rice orthologue of the gene containing the most significantly associated barley SNP. However,

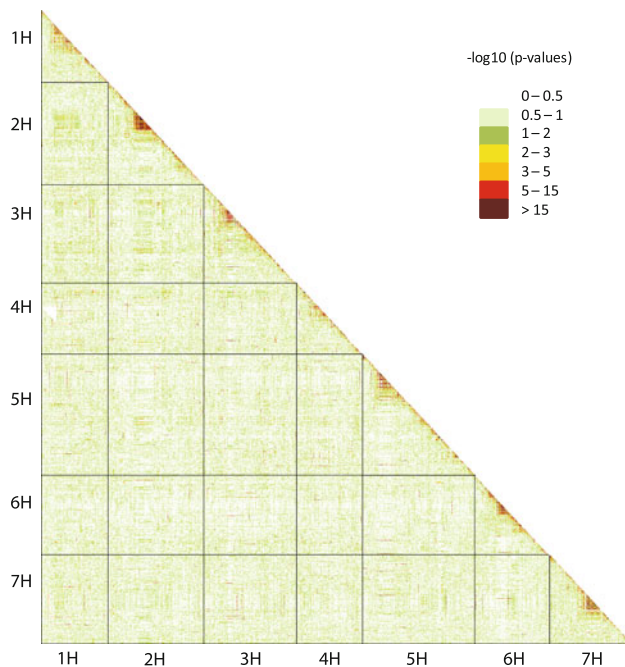


Fig. 3 Genome-wide association scan for 2,132 mapped SNPs. $-\log_{10}(f_p \text{ value})$ scores for all SNP-SNP pairwise comparisons accounting for relatedness (kinship) are shown

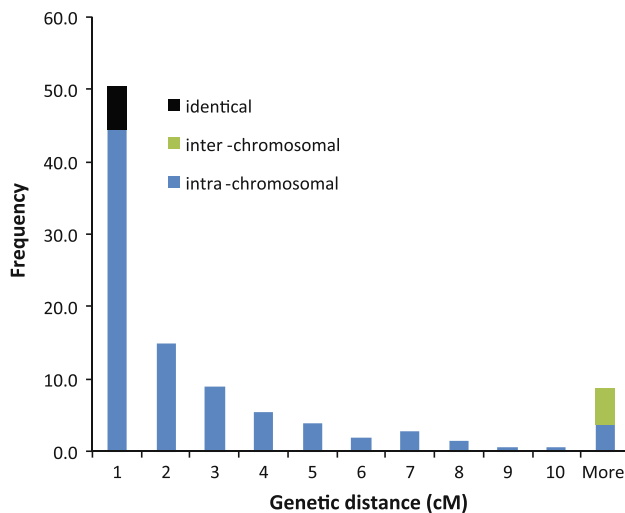


Fig. 4 Mapping resolution. Genome-wide association scans for 2,132 mapped SNPs were performed accounting for relatedness (kinship). Most significant marker genetic distances to the known map location of the mapped SNP were used to evaluate the mapping resolution attained and amount of putative false positives. Percentages of SNP markers mapping within 10 cM, and further than 10 cM are shown

mapping resolution (based on rice gene model estimates) rapidly decreased when approaching the genetic centromeres: 25% of the SNPs immediately flanking the centromeric regions exhibit a resolution approaching 190 gene models, with those located in the genetic centromeres exhibiting a mapping resolution ranging from 200 to over

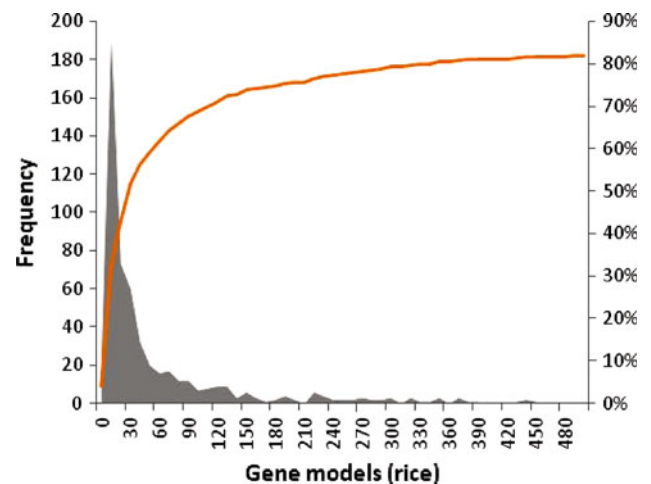


Fig. 5 Mapping resolution. Mapping resolution expressed in number of gene models gene content inferred from the rice genome annotation site <http://rice.plantbiology.msu.edu/>. Most of the remaining 18% (123 SNPs) not shown in the graph are centromeric and intervals delimited from 500 to 3,000 gene models

1,000 gene models. Thus, the gene-rich but rarely recombining haplotype blocks extending across the genetic centromeres, which cannot be resolved in bi-parental mapping populations, still cannot be resolved in genetically diverse association mapping populations of comparable size.

Discussion

Barley has a large 5,300 Mb un-sequenced genome but extensive EST resources derived from nine cultivated lines (Harvest, Barley v1.68) (Wanamaker et al. 2008). Using this EST information Close et al. (2009) previously developed three Illumina 1,536-plex gene-based SNP assay platforms from a combination of informatics analysis and by re-sequencing PCR-amplicons from a collection of eight diverse elite barley cultivars. They used these POPA's to genotype three-doubled haploid barley mapping populations [Steptoe \times Morex (Kleinohs et al. 1993), Morex \times Barke (Stein et al. unpublished) and OWB-D \times OWB-R (Costa et al. 2001)] and generated genetic linkage maps of each population. Then, they used a directed acyclic graphing algorithm implemented in MergeMap (Wu et al. 2008) to derive a consensus map from the forced linear order of the 2,943 polymorphic SNPs segregating in the three populations. The consensus map coordinates from MergeMap were normalized to the arithmetic mean cM distance for each linkage group from the individual maps. We considered this consensus map to represent an approximate gene order along each of the seven barley chromosomes and used this as a template for GWAS. We chose to remove rare SNPs from our GWAS datasets.

While this is common practice, it results in a huge loss of information and limits our ability to capture variation associated with rare alleles. Loci with a low MAF (<10%) have less power to detect weak genetic effects than loci with a high MAF (>40%) because of small sample size (Ardlie et al. 2002). Furthermore, previous studies have demonstrated that rare genotypes are more likely to result in spurious findings (Lam et al. 2007) because of a higher relatedness between individuals sharing rare alleles. While it has been shown, in large human GWAS, that including SNP loci with MAF > 5% does not result in inflated false positive rates (Tabangin et al. 2007), due to the complexity of the pedigrees linked to plant populations we decided to remove SNPs with MAF < 10% from our LD and GWAS.

The patterns of genetic diversity along each of the seven barley genetic linkage maps varied amongst the sub-groups identified by both PCO and STRUCTURE analysis. However, the overall genetic diversity in the population remained high. We did observe a 2.9 cM region on barley chromosome 3H that exhibited a sharp decrease in genetic diversity across all germplasm groups. This interval would contain 585 gene models if we assumed absolute conservation of synteny between rice and barley. It may represent a strong signature of selection for non-brittle rachis, a trait involved in non-shattering of ears after ripening and that was important in barley domestication (Komatsuda and Mano 2002; Komatsuda et al. 2004). The position of this 2.9 cM interval on the short arm of chromosome 3H is consistent with that reported in previous studies for non-brittle rachis loci (Kandemir et al. 2004). The BCD706 and ABG396 RFLP markers delimiting the brittle rachis QTL interval (Kandemir et al. 2004) co-segregate with BOPA markers 11_10081, and 11_10137, respectively (Szucs et al. 2009) delimiting a 14.95 cM interval on the consensus map of Close et al. (2009). Brittle rachis in wild barley is controlled by two dominant complementary genes, *Btr1* and *Btr2*, with mutations in either locus (*btr1* or *btr2*) resulting in the non-brittle rachis of cultivated barley. The *btr1* allele is present in most occidental cultivars whereas the *btr2* allele is present in most oriental cultivars. Interestingly, we did not observe differential patterns of diversity in this region between the European and American Manchurian types. Only seven lines were polymorphic in the region: Mehola and OWB-R which are both brittle rachis lines and Dicktoo, Haruna Hijo, Morex, Steptoe and OWB-D, which count among the genetically “exotic” cultivars used as the parental lines of several mapping populations.

The extent of LD has long been of interest for population geneticists as its value determines the required genetic marker number and mapping resolution achievable in GWAS. It is commonly accepted that the extent of LD over short genetic distances is mainly affected by recombination while population structure (or possibly epistasis) largely

accounts for long-range LD. Classic algorithms to measure LD (r^2/D') are useful to explore short and long-range patterns of LD. However, they fail to discriminate between LD caused by genetic linkage and that caused by population structure (Text S1). We therefore also used a mixed model approach with relative kinship estimators as co-factors to investigate LD patterns without most of the population structure effects (Yu et al. 2006). These results can be extrapolated to indicate the expected frequency of false positives and expected resolution in subsequent whole genome scans when using the same marker dataset and statistical model. While we observed high LD extending along the spine of each chromosome, as expected, the mixed model significantly reduced the long-range and inter-chromosomal (background) LD and we subsequently used this approach to assess the mapping resolution achievable by GWAS.

91.2% of SNPs were positioned by GWAS to within 10 cM of their position predicted on the consensus map of Close et al. (2009). Inspection of the patterns of LD for the 8.8% of SNP markers that did not map within 10 cM of their original genetic map position revealed 3.7 and 5.1% intra- and inter-chromosomal SNP:SNP associations, respectively. We did not find any that mapped into the centromeric regions. This set of markers and those that mapped within 5–10 cM of their original positions are useful for identifying genomic regions where current marker coverage is insufficient, either as a result of very fast decay of LD or the presence of non-tagged SNPs (which we define as those that segregate in only a subset of the germplasm and are not in LD with their flanking SNPs, despite being physically close). The three genetic clusters observed within our sample (two-row spring barleys, winter barleys and Manchurian types) have been in distinct breeding pools for a considerable period of time (Malysheva-Otto et al. 2006) and only a few individuals resulting from inter-cluster crosses, which could potentially introduce recombination events between clusters, are present in the sample. Thus, there is the possibility that SNPs present in only one cluster are not in LD with flanking SNPs at the whole sample level. Supporting this hypothesis we did observe contrasting PIC values in closely linked SNPs. Further investigation of the SNPs that did not map close to their original genome position and the SNPs that map to more than one genome position should be pursued to investigate the fraction of those related directly to genetic map artifacts, spurious associations due to population structure and those related to gene duplication events and epistasis.

Our results have important implications for the design of association mapping studies in barley. We have shown that genetically diverse populations of relatively small size prove adequate for fine-mapping simple traits, as long as the trait is segregating across the entire mapping population

(MAF > 10% in our case), that population sub-structure is effectively controlled in the analysis and the trait does not fall into recombinationally poor regions such as genetic centromeres. Despite these limitations, we show that 87.3% of the SNPs could be mapped by GWAS to within 5 cM (50% within 1 cM) of their position on the Close et al. (2009) consensus map. We also show that if the positive associations do not fall within centromeric regions the mapping resolution achieved was reasonably high (i.e. 50% of the test SNPs mapped to within 27 gene models of the GWAS framework markers). In practical terms, this means that SNPs associated to a simple trait or quantitative locus with strong additive effects are potentially close enough to be used with reasonable confidence in marker-assisted breeding.

The use of markers based on gene sequence data is of special interest because they facilitate exploitation of conservation of synteny with model, fully sequenced grass genomes (e.g. *Brachypodium* and rice). Consequently, considerable value can be attributed to the targeted identification of new markers that are even closer to a positive association, allowing the interval containing the causal gene to be better delimited. Conserved synteny also helps to predict the number and identity of possible candidate genes, forming the focus of further investigations that may include allele re-sequencing across the association panel to improve genetic resolution, screening an independent GWA panel containing different germplasm or, when available, re-sequencing an allelic series of mutants that affect the target trait. Clearly, the situation is different for traits where candidate genes are not obvious, where there is a breakdown in the conservation of synteny and/or where well-characterized mutant stocks are not available. In those cases, identification of the causal genes will most likely proceed in combination with large and relevant bi-parental mapping populations that can be used for validation and, if considering quantitative characters, after the generation of QTL-near isogenic lines containing alternative alleles. Here, a barley genome sequence will have a big role to play, cutting out the ‘middle men’ (rice/*Brachypodium*), and providing a true list of positional candidates for more detailed investigation (Schulte et al. 2009).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 71:304–311
- Badr A, Muller K, Schafer-Pregl R, El Rabey H, Effgen S et al (2000) On the origin and domestication history of barley (*Hordeum vulgare*). *Mol Biol Evol* 17:499–510
- Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331
- Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172:557–567
- Chandler VL, Brendel V (2002) The maize genome sequencing project. *Plant Physiol* 130:1594–1597
- Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N et al (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10:582
- Costa JM, Corey A, Hayes PM, Jobet C, Kleinhofs A et al (2001) Molecular mapping of the Oregon Wolfe Barleys: a phenotypically polymorphic doubled-haploid population. *Theor Appl Genet* 103:415–424
- Ersoz ES, Yu J, Buckler ES (2008) Applications of linkage disequilibrium and association mapping in crop plants. In: Varshney RK, Tuberosa R (eds) *Genomic assisted crop improvement: vol I: Genomic approaches, platforms*. Springer Verlag, Germany
- Ersoz ES, Yu J, Buckler ES (2009) Applications of linkage disequilibrium and association mapping in maize *Molecular Genetic Approaches to Maize Improvement*, Springer Berlin Heidelberg, pp 173–195
- FAO (2002) World agriculture: towards 2015/2030. Summary report. Rome, Food and Agriculture Organization of the United Nations
- Felsenstein J (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst Biol* 46:101–111
- Goff SA, Ricke D, Lan TH, Presting G, Wang R et al (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. japonica). *Science* 296:92–100
- Hudson RR, Slatkin M, Maddison P (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics* 132:583–589
- Jun Yu, Songnian Hu, Wang Jun, Gane Ka-Shu Wong, Li Songgang et al (2002) A Draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 296:79–92
- Kandemir N, Yildirim A, Kudrna DA, Hayes PM, Kleinhofs A (2004) Marker assisted genetic analysis of non-brittle rachis trait in barley. *Hereditas* 141:272–277
- Kleinhofs A, Kilian A, Saghai Maroof MA, Biyashev RM, Hayes PM et al (1993) A molecular, isozyme and morphological map of the barley (*Hordeum vulgare*) genome. *Theor Appl Genet* 86:705–712
- Komatsuda T, Mano Y (2002) Molecular mapping of the intermedium spike-c (int-c) and non-brittle rachis 1 (btr1) loci in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 105:85–90
- Komatsuda T, Maxim P, Senthil N, Mano Y (2004) High-density AFLP map of nonbrittle rachis 1 (btr1) and 2 (btr2) genes in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 109:986–995
- Kruglyak L (2008) The road to genome-wide association studies. *Nat Review Genet* 9:314–318
- Lam AC, Schouten M, Aulchenko YS, Haley CS, de Koning DJ (2007) Rapid and robust association mapping of expression quantitative trait loci. *BMC Proc* 1(Suppl 1):S144
- Lawrence CJ, Harper LC, Shaeffer ML, Sen TZ, Seigfried TE et al (2008) MaizeGDB: the maize model organism database for basic, translational, and applied research. *Int J Plant Genomics*
- Malysheva-Otto LV, Ganai MW, Roder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genet* 24:6–7

- Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2008) GenStat for Windows Introduction, 11th edn. VSN International, Hemel Hempstead, UK
- Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 60:227–237
- Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. *Am J Hum Genet* 67:170–181
- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR et al (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci USA* 103:18656–18661
- Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T et al (2009) The international barley sequencing consortium—at the threshold of efficient access to the barley genome. *Plant Physiol* 149:142–147
- Szucs P, Blake VC, Bhat PR, Chao S, Close TJ et al (2009) An integrated resource for barley linkage map and malting quality QTL alignment. *The Plant Genome* 2:134–140
- Tabangin ME, Woo JG, Liu C, Nick TG, Martin LJ (2007) Comparison of false-discovery rate for genome-wide and fine mapping regions. *BMC Proc* 1(Suppl 1):S148
- The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Wanamaker SI, Close TJ, Roose ML, Lyon M (2008) HarvEST <http://harvest.ucr.edu>
- Waugh R, Jannink JL, Muehlbauer GJ, Ramsay L (2009) The emergence of whole genome association scans in barley. *Curr Opin Plant Biol* 12:218–222
- Wu Y, Close TJ, Lonardi S (2008) On the accurate construction of consensus genetic maps. *Comput Syst Bioinformatics Conf* 7:285–296
- Yu J, Pressoir G, Briggs WH, Vroh B, Yamasaki IM et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zamir D (2008) Plant breeders go back to nature. *Nat Genet* 40:269–270
- Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. *The Plant Genome* 1:5–20